Deep Scattering Power Spectrum Features for Robust Speech Recognition

INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Deep Scattering Power Spectrum Features for Robust Speech Recognition

Neethu M. Joy [1], *Dino Oglic* [1], *Zoran Cvetkovic* [1], *Peter Bell* [2], *and Steve Renals* [2]

[1] Department of Engineering, King's College London, UK
[2] Center for Speech Technology Research, University of Edinburgh, UK
{neethu.joy, dino.oglic, zoran.cvetkovic}@kcl.ac.uk, {peter.bell, s.renals}@ed.ac.uk

Abstract

Deep scattering spectrum consists of a cascade of wavelet transforms and modulus non-linearity. It generates features of different orders, with the first order coefficients approximately equal to the Mel-frequency cepstrum, and higher order coefficients recovering information lost at lower levels. We investigate the effect of including the information recovered by higher order coefficients on the robustness of speech recognition. To that end, we also propose a modification to the original scattering transform tailored for noisy speech. In particular, instead of the modulus non-linearity we opt to work with power coefficients and, therefore, use the squared modulus non-linearity. We quantify the robustness of scattering features using the word error rates of acoustic models trained on clean speech and evaluated using sets of utterances corrupted with different noise types. Our empirical results show that the second order scattering power spectrum coefficients capture invariants relevant for noise robustness and that this additional information improves generalization to unseen noise conditions (almost 20% relative error reduction on AURORA4). This finding can have important consequences on speech recognition systems that typically discard the second order information and keep only the first order features (known for emulating MFCC and FBANK values) when representing speech.
Index Terms: scattering coefficients, wavelet transform, robustness, deep scattering spectrum, power spectrum

1. Introduction

Speech recognition systems typically operate in feature spaces designed to implement invariances inherent to speech production and human speech recognition [1, 2, 3]. Log Mel-filter bank values (FBANK) and Mel-frequency cepstral coefficients (MFCC) are two feature extraction techniques known for effectively modeling local invariances at short time scales of 25 ms [1, 4, 5]. As established in [6], fundamental for the effectiveness of these two techniques is the approximate Lipschitz continuity of the feature extraction operator [7]. FBANK and MFCC features achieve this by performing weighted power spectra averaging. Whilst power spectra averaging is important for the approximate Lipschitz continuity, the operator at the same time performs compression that can result in information loss [8, 9]. To mitigate that, Mallat [6] has introduced the scattering operator that computes coefficients using a *cascade of wavelet transform*s and *modulus non-linearity*. The first order scattering coefficients are designed to emulate FBANK features and the higher order ones to recover the information lost due to the waveform compression in coefficients of lower orders. In [6] it has been hypothesized that higher order coefficients capture invariants relevant to amplitude modulations lost in the first order scatter, as well as frequency transpositions required for speaker-independent classification of phonetic units.

Whilst it is true that higher order coefficients can recover information lost due to compression in lower levels of cascade, there has not been a comprehensive empirical evidence for the relevance of recovered information to robustness of speech recognition systems. In particular, Sainath et al. [10] have investigated the relevance of second order information on news recordings that can have various competing acoustic sources but are not considered to be particularly noisy. In that study, it has been established that second order information amounts to 4-7% relative improvement across different settings (multi-resolution time and frequency scatter, data adaptation, and sequence training). Moreover, it was empirically demonstrated that on that particular dataset the same improvement can be obtained by employing multi-resolution FBANK and MFCC features. Similarly, Fousek et al. [11] have investigated the relevance of second order information on IBM voice search data using multi-layer perceptrons. The main finding was that second order information contributes to 2-4% of relative improvement in accuracy, depending on the employed normalization scheme. We extend these two studies and investigate the merits of second order information in the context of noisy speech. Our focus is on the setting with a significant mismatch between training and test sets which can provide a good estimate of generalization abilities to unseen noise environments. In particular, we train our acoustic models in clean conditions and evaluate them using sets of utterances corrupted with different noise types. This is different from a typical regime for learning robust acoustic models known as *multi-condition training* in which the clean set of training recordings is augmented with noise corrupted utterances. The latter introduces confounding effects into the training process and the test error of such models might not be a good estimate for the ability of higher order scattering coefficients to capture noise robust invariants and generalize to unseen noise types.

Our empirical findings rely on a technical modification to the original scattering operator. In particular, instead of the modulus non-linearity proposed in [6] we opt to work with power coefficients and, therefore, use the *squared modulus non-linearity*. This choice was motivated by previous research showing that it can result in removal of spurious noise components from scattering features by means of utterance level normalizations [12, 13, 14]. To empirically establish the relevance of information in higher order scattering coefficients on generalization abilities of acoustic models, we perform a series of experiments on AURORA4 [15], common and still challenging benchmark for noise robust speech recognition. An advantage of this dataset is that it is possible to evaluate merits of a feature representation without confounding effects typically introduced by means of data augmentation and multi-condition training. In particular, in the first set of experiments we evaluate the merits of information in the second order coefficients by training on clean speech and evaluating on noisy utterances with a significant mismatch between training and test conditions. In this setting, we show that scattering power spectrum features provide a robust representation of waveform frames and that second order information

http://dx.doi.org/10.21437/Interspeech.2020-2656

amounts to almost 20% relative improvement in the accuracy. In the second experiment, we evaluate the approach using multi-condition training where we establish that the performance gap is reduced as a result of convolutional blocks being able to capture relevant patterns from first-order coefficients. An open question is whether such patterns would generalize to completely novel noise environments and whether they account for spurious correlations introduced by data augmentation characteristic to multi-condition training. Despite the reduction in performance gap, the information in second order coefficients still amounts to 12% relative improvement in the accuracy. Moreover, in this training regime we observe a relative improvement of 14% compared to multi-resolution FBANK features which is contrary to findings in [10] on news data. In our third experiment, we evaluate the effectiveness of multi-resolution scattering cascades and show that such networks perform on par with state-of-the-art deep convolutional models with more complex architectures.

2. Deep Scattering Power Spectrum

Given a time-domain signal $x(t)$, the first order scattering coefficients are generated according to [6]

$$S_1(t, \lambda_1) = |x * \psi_{\lambda_1}| * \phi(t) \ , \tag{1}$$

where $\psi_{\lambda_1}(t), \lambda_1 \in \Lambda_1$, is a bank of band-pass filters obtained from a mother wavelet $\psi(t)$ via time scaling specified by factors λ_1, and $\phi(t)$ is a low-pass filter that performs local averaging. The scaling parameters λ_1 are distributed uniformly on the logarithmic scale, and the number of filters per octave, Q_1, sets the frequency resolution of the transform. The first order scattering coefficients approximate the Mel-frequency spectrum [5], and if the octave resolution is set to $Q_1 = 8$, then the wavelet filters have the same frequency resolution as a Mel-filterbank (in total, 41 filters). The low-pass filter $\phi(t)$, which is a Hamming window with a time support of 25 ms, ensures that the scattering representation is locally invariant to time shifts smaller than 25 ms. Just as in FBANK and MFCC features, for applications in speech recognition the logarithm is applied to scattering coefficients to mimic psychoacoustic measurements and physiology of human hearing [2, 3, 4, 5]. The resulting feature vector is referred to as the *first order time scatter* (illustration in Figure 1).

While weighted averaging by $\phi(t)$ provides locally translation invariant and distortion stable features, it at the same time results in loss of information regarding transient phenomena and finer amplitude modulations in speech signals. The information lost in $S_1(t, \lambda_1)$ can be recovered by processing the sub-band signals $\{|x * \psi_{\lambda_1}(t)|\}_{\lambda_1 \in \Lambda_1}$ using another constant Q wavelet filter bank $\{\psi_{\lambda_2}(t)\}_{\lambda_2 \in \Lambda_2}$, as illustrated in Figure 1. Typically, the resolution of Λ_2 is set to $Q_2 = 1$ in order to get a sparse representation, concentrating signal information over as few wavelet coefficients as possible. The wavelets in Λ_2 have a narrow time support and are better adapted to characterize transients and attacks. The second order scattering operator is given by

$$S_2(t, \lambda_1, \lambda_2) = ||x * \psi_{\lambda_1}| * \psi_{\lambda_2}| * \phi(t) \ , \tag{2}$$

with $\lambda_1 \in \Lambda_1$ and $\lambda_2 \in \Lambda_2$. The second order scattering coefficients are typically normalized by dividing by the corresponding first order coefficient, i.e., $S_2(t, \lambda_1, \lambda_2) / S_1(t, \lambda_1)$. Henceforth, the logarithm of the normalized second order coefficient will be referred to as the *second order time scatter*. Although this cascade of wavelet transforms can be further extended to higher order coefficients, in [5] it has been demonstrated that for $\phi(t)$ with support of 25 ms, almost 99.3% of signal energy can be

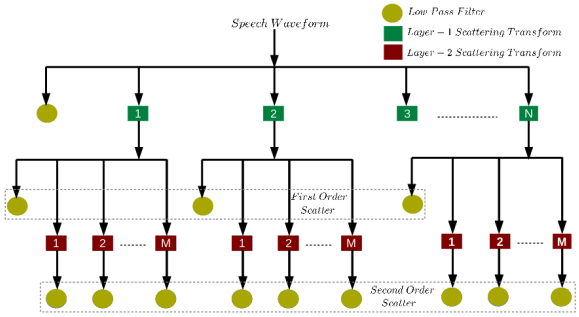


Figure 1: *The figure illustrates a cascade of wavelet transforms for extracting first and second order time-scattering coefficients.*

recovered with S_1 and S_2 operators. Hence, in our experiments, we have limited the scattering order to 2. Further invariances to speaker-specific frequency transpositions can be achieved by applying the scattering operations to $S_1(t, \lambda_1)$ and $S_2(t, \lambda_1, \lambda_2)$ along their λ variables, which is referred to as *frequency scatter*. Combining features obtained through the frequency scatter operator with the first and second order time scatter has shown to improve the performance on LVCSR tasks [10]. However, results of our experiments focused on robustness of automatic speech recognition were inconclusive with regard to merits of the frequency scatter, and in fact the best results were obtained with the first and second order time scatter only, so henceforth the frequency scatter is not considered.

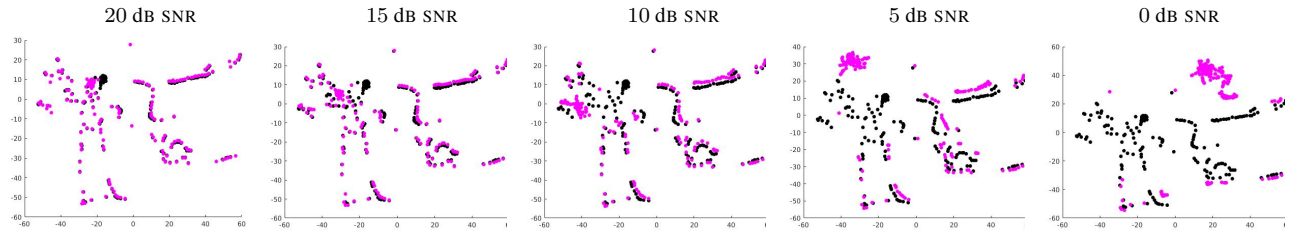
The modulus non-linearity was selected by Mallat [6] to suppress the effects of large scattering coefficients. However, in noisy settings we would like to do the opposite and amplify large wavelet coefficients, thereby suppressing spurious noise components. Thus, to make the scattering transform more robust to noise, we propose to replace the original non-linearity with modulus squared. More formally, the first and second order scattering transforms from Eq. (1) and (2) are now computed by

$$\hat{S}_1(t, \lambda_1) = |x * \psi_{\lambda_1}|^2 * \phi(t) \ , \tag{3}$$

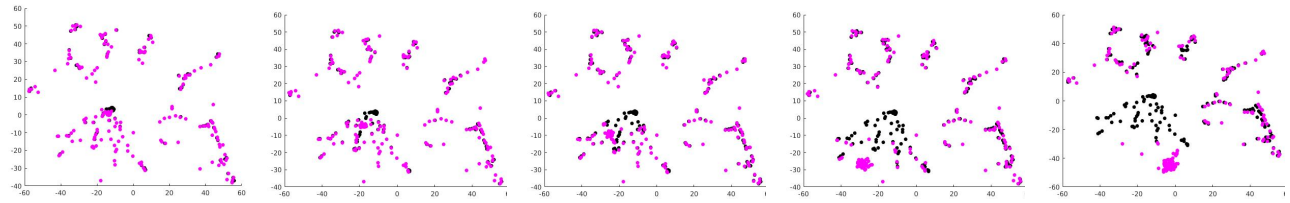
$$\hat{S}_2(t, \lambda_1, \lambda_2) = \left||x * \psi_{\lambda_1}|^2 * \psi_{\lambda_2}\right|^2 * \phi(t) \ , \tag{4}$$

with $\lambda_1 \in \Lambda_1$ and $\lambda_2 \in \Lambda_2$. This choice of non-linearity can also be motivated by previous work in signal processing that aims at removing spurious noise contributions by means of utterance level normalizations [12, 13, 14], which we will also employ in its simplest form in our experiments. We observed that another benefit of the squared modulus non-linearity is in that it makes the feature representation much sparser than conventional deep scattering spectrum. Henceforth, we refer to the logarithm of first order coefficients, computed according to $\hat{S}_1(t, \lambda_1)$ with $\lambda_1 \in \Lambda_1$, as the *first order scattering power spectrum* coefficients. The *second order scattering power spectrum* coefficients are obtained by applying the logarithm operator to normalized coefficients, $\hat{S}_2(t, \lambda_1, \lambda_2) / \hat{S}_1(t, \lambda_1)$ with $\lambda_1 \in \Lambda_1$ and $\lambda_2 \in \Lambda_2$.

In Figure 2, we illustrate the difference in capturing invariants relevant for noise robustness between the two non-linearities. In particular, we take a clean utterance from TIMIT [16] (*she had your dark suit in greasy wash water all year*) and compute scattering representations of waveform frames using both non-linearities, modulus characteristic to DSS and squared modulus proposed here. Following this, we corrupt the original utterance using additive Gaussian noise with SNR varying from 0-20 dB. The scattering representations with the two non-linearities are then computed for the noise corrupted utterances. To visualize the manifold where the data lies we use the t-SNE toolkit [17].



(a) Deep scattering spectrum features for a clean utterance and its white noise added counterpart at various SNR *levels.*



(b) Deep scattering power spectrum features for a clean utterance and its white noise added counterpart at various SNR *levels.*

Figure 2: *Deep scattering power spectrum vs standard* DSS *representation [6] of a clean utterance and its noise corrupted counterparts, with* SNR *levels varying from* 0-20 *d*B. *The black points represent clean speech frames and the purple ones their noisy counterparts.*

We can see from the figure that the DSS representation starts to degrade at about 5 dB, measured with the increased spread between the positioning of clean and noisy frames. In contrast to this, scattering power spectrum features that employ squared modulus non-linearity tend to keep clean and corresponding noisy utterance within balls of small radius. As a result, it is likely that labels assigned to such frames will be identical because the learned hypothesis are typically smooth in small neighborhoods around a particular training point in the instance space. It is promising that the squared modulus non-linearity keeps clean and noisy frames quite close to each other even at 0 dB SNR.

3. Experiments

We perform our experiments on AURORA4 [15] dataset, a standard benchmark for noisy speech. This is a medium vocabulary task based on a clean speech *training set* from the Wall Street Journal (WSJ0) corpus [18]. The clean speech was corrupted by six different noise types (street traffic, train station, car, babble, restaurant, airport) at 10-20 dB SNR. The dataset allows for training using two different sets of observation: *i*) clean condition mode that accounts only for clean speech utterances, and *ii*) multi-condition mode that contains a mix of clean and noise corrupted utterances. The evaluation set is derived from WSJ0 5K-word closed-vocabulary test set, consisting of 330 utterances from 8 speakers. These test utterances were recorded by a primary and a secondary microphone. Each of the two *test sets* are then corrupted by the same noise types used in multi-condition training set but this time at 5-15 dB SNR (in total 14 test sets).

Scattering features for both considered non-linearities (modulus and squared modulus) were extracted using the SCATNET [1] toolkit. If not specified otherwise, we are generating first order coefficients using Gabor filters with resolution $Q_1 = 8$, which gives in total 41 features. The second order coefficients are generated using Morlet wavelets with resolution $Q_2 = 1$, which for $Q_1 = 8$ gives in total 127 features (coefficients with small values are truncated). The frame context is always set to 5 so that in total 11 successive frames are stacked and passed to the network as input. We have investigated the effects of mean normalization on both speaker and utterance levels for these coefficients and established that the latter is more effective in noisy settings. For experiments with FBANK features, we first apply pre-emphasis and then split the signal into 25 ms long frames (with 10 ms shifts) using the Hamming window. If not specified otherwise, we use 40 FBANK features which are normalized at the utterance level by subtracting the corresponding mean.

[1] https://www.di.ens.fr/data/software/scatnet

3.1. Network Architecture

The main challenge in incorporating second order information into speech recognition systems is in that first order coefficients typically require some further band-pass filtering to capture local invariants in such spectro-temporal decompositions of speech waveforms. Empirically, the most effective neural architecture for hybrid acoustic models with scattering features has been proposed in [19]. The architecture is a *junction network* that takes as input first and second order scattering coefficients via separate pipelines which are then merged into a multi-layer perceptron with several hidden layers. While the first pipeline is a convolutional network that takes first order coefficients as inputs, the second one is a multi-layer perceptron with a single hidden layer that extracts features from second order coefficients.

The convolutional pipeline is comprised of 3 layers, each consisting of the following blocks: *i*) one dimensional convolution, *ii*) max-pooling, *iii*) layer normalization [20], and *iv*) RELU nonlinearity followed by dropout [21]. In the first layer, we use convolutions with 80 channels and filter size 10 samples. The second and third layer, on the other hand, use 60 channels and filter size 3 samples. The pooling is applied with compression rates 3, 2 and 1, respectively. We note here that our configuration of channels and filter sizes is different from prior work [10, 19]. The pipeline for second order coefficients is a multi-layer perceptron (MLP) with 512 activation units, followed by batch normalization [22], RELU non-linearity, and a dropout block. The outputs of the two pipelines are merged into a vector and then passed to MLP with 6 hidden layers, each having 1024 activation units. We use batch normalization after each dense layer in MLP, followed by RELU non-linearity and a dropout block. The dropout rate across the network is set to 0.15.

3.2. Clean Condition Training

In the first experiment, the goal is to establish that second order scattering coefficients contain invariants relevant for robust-

FEATURES	A_1	B_{2-7}	C_8	D_{9-14}	AVG_{1-14}
$DSPS_1$	2.76	13.83	7.74	17.90	14.35
$DSPS_1$ + $DSPS_2$	2.58	**11.14**	**6.89**	**14.33**	**11.59**
DSS_1 [5]	2.62	14.72	7.89	19.07	15.23
DSS_1 + DSS_2 [5]	2.61	11.95	7.33	15.33	12.40
$FBANK_{40}$ [4]	2.65	13.75	7.96	16.89	13.89
$FBANK_{60}$ [4]	2.54	13.06	8.33	17.08	13.69
$FBANK_{80}$ [4]	2.69	12.04	8.03	16.19	12.86
$FBANK_{100}$ [4]	**2.52**	12.60	7.60	16.52	13.20

Table 1: *The table reports word error rates obtained on test sets (*14 *in total) of* AURORA4 *by* ***clean condition*** *training. First order features (including* FBANK*s) are processed using a convolutional neural network, while the combination of first and second order features employs the junction architecture from [19].*

FEATURES	A_1	B_{2-7}	C_8	D_{9-14}	AVG_{1-14}
$DSPS_1$	2.97	5.88	6.71	15.96	10.05
$DSPS_1$ + $DSPS_2$	2.73	**5.20**	**4.73**	**14.15**	**8.83**
DSS_1 [5]	2.99	5.69	6.56	15.95	9.96
DSS_1 + DSS_2 [5]	2.86	5.45	6.11	15.08	9.44
$FBANK_{40}$ [4]	3.06	6.08	7.10	16.09	10.23
$FBANK_{60}$ [4]	2.90	5.72	6.46	15.65	9.83
$FBANK_{80}$ [4]	2.88	5.58	5.92	15.22	9.55
$FBANK_{100}$ [4]	**2.69**	5.33	5.74	15.26	9.43

Table 2: *The table reports word error rates obtained on test sets (*14 *in total) of* AURORA4 *by* ***multi-condition*** *training. First order features (including* FBANK*s) are processed using a convolutional neural network, while the combination of first and second order features employs the junction architecture from [19].*

ness and that they generalize to unseen noise environments. To demonstrate this, we train using clean speech and evaluate our model on the noise corrupted test sets with a significant mismatch between training and test conditions. To quantify the improvement that comes as a result of including the information from second order coefficients, we train first with the convolutional pipeline for processing the first order coefficients (with second order pipeline switched off). We refer to this model as $DSPS_1$ to express the fact that its inputs are first order deep scattering power spectrum (DSPS) coefficients. Following this, we train the junction network jointly, thereby accounting for both first and second order scattering coefficients. This model is referred to as $DSPS_1$ + $DSPS_2$ to account for the fact that the architecture takes both first and second order coefficients as inputs. We use the relative improvement in the accuracy between these two models to quantify the relevance of the information from second order coefficients for noise robustness. Table 1 provides a summary of our experiments in clean conditions. The empirical evidence indicates a relative improvement of almost 20% with the addition of second order information, compared to convolutional network with first order features ($DSPS_1$ vs $DSPS_1$ + $DSPS_2$, utterance normalization). In the same setting, we also compare to deep scattering spectrum that employs modulus non-linearity when generating features. The experiments, along with the illustration in Figure 2, demonstrate that square modulus non-linearity and power spectrum scattering provide a more robust representation than conventional DSS, achieving over 6% relative improvement. In addition to all of this, we run an experiment with FBANK features aimed at showing that invariants recovered by second order coefficients cannot be obtained by increasing the granularity of first order features. For that, we train the convolutional pipeline/network with different number of FBANK features and show that the junction network combining first and second order information outperforms all such models, including 16.5% relative improvement over FBANK features with the same resolution.

3.3. Multi-Condition Training

In the second experiment, we first train the same set of models as in clean condition training. Table 2 provides a summary of our empirical results in this setting. The results indicate that again learning with scattering power spectrum features is more effective than with conventional DSS, achieving similar relative improvement of over 6%. Moreover, the experiments with FBANK features also indicate that one cannot recover invariants from second order coefficients by just increasing the number of filters in the first order coefficients. An interesting observation is that the gap between first order pipeline and the junction network is decreased compared to clean condition training. Still, the information in second order coefficients amounts to 12% relative improvement in the accuracy. Our hypothesis is that the convolutional blocks are in this setting capable of capturing patterns from the first order coefficients that are relevant for robustness. However, it is unclear whether the captured patterns account for spurious correlations introduced by data augmentation characteristic to multi-condition training nor whether they would generalize to unseen noise environments.

In our final experiment, we investigate the effectiveness of models that combine features with multiple resolutions and use multi-condition training. Table 3 provides a summary of our results in this setting. We can observe that combination of resolutions $Q = \{4, 13\}$ provides the lowest error rate over test samples. Moreover, that model outperforms state-of-the-art deep convolutional networks with 6 and 10 such layers that take FBANK features as inputs [23]. Interestingly, the relatively small junction network consisting of only three convolutional layers, supplemented with information from second order coefficients, is competitive with a recently proposed architecture with 15 layers of much more expressive multi-octave convolutions [24].

ARCHITECTURE	CNN DEPTH	AVG_{1-14}
$DSPS_1$ + $DSPS_2$ (MULTI-RESOLUTION SCATTERING)		
$Q = \{8\}$	3	8.83
$Q = \{1, 4, 13\}$	3	8.76
$Q = \{1, 4, 8, 13\}$	3	8.94
$Q = \{4, 13\}$	3	**8.64**
FBANK BASELINES		
FMLLR + MLP	-	10.21
VD6CNN [23]	6	10.34
VD10CNN [23]	10	8.81
M-OCT CNN [24]	15	**8.31**

Table 3: *A summary of results obtained on* AURORA4 *by combination of multi-resolution scatters and multi-condition training.*

4. Conclusion

We have proposed a modification to the scattering transform that is capable of recovering information lost due to compression by first order features such as FBANK and MFCC coefficients. In our empirical analysis, we have demonstrated that: *i*) the first order features discard information relevant for noise robust speech recognition, and *ii*) the second order deep scattering power spectrum coefficients capture invariants relevant for noise robust speech recognition and that these invariants can generalize to noise types not contained in the training set. Moreover, our empirical results suggest that second order coefficients can lead to gross simplification of neural architectures and, thus, reduce training time and the amount of required computing resources.

5. Acknowledgements

This work was supported in part by EPSRC grant EP/R012067/1 (SPEECHWAVE). The authors would also like to thank Erfan Loweimi for constructive feedback and help with generating Kaldi alignments.

6. References

[1] J. S. Bridle and M. Brown, "An experimental automatic word-recognition system," JSRU, Ruislip, UK, Tech. Rep. 1003, 1974.

[2] R. C. Moore, T. Lee, and F. E. Theunissen, "Noise-invariant neurons in the avian auditory cortex: Hearing the song in noise," *PLOS Computational Biology*, vol. 9, no. 3, pp. 1–14, 2013.

[3] F. Li, A. Trevino, A. Menon, and J. Allen, "A psychoacoustic method for studying the necessary and sufficient perceptual cues of american english fricative consonants in noise," *The Journal of the Acoustical Society of America*, vol. 132, pp. 2663–75, 2012.

[4] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, 1980.

[5] J. Andén and S. Mallat, "Deep scattering spectrum," *IEEE Transactions on Signal Processing*, 2014.

[6] S. Mallat, "Group invariant scattering," *Communications on Pure and Applied Mathematics*, 2012.

[7] H. Gouk, E. Frank, B. Pfahringer, and M. Cree, "Regularisation of neural networks by enforcing Lipschitz continuity," *arXiv:1804.04368*, 2018.

[8] J. Yousafzai, P. Sollich, Z. Cvetkovic, and B. Yu, "Combined features and kernel design for noise robust phoneme classification using support vector machines," *IEEE Trans. Audio, Speech and Lang. Proc.*, vol. 19, pp. 1396–1407, 2011.

[9] M. Ager, Z. Cvetkovic, and P. Sollich, "Combined waveform-cepstral representation for robust speech recognition," *IEEE ISIT*, 2011.

[10] T. N. Sainath, V. Peddinti, B. Kingsbury, P. Fousek, B. Ramabhadran, and D. Nahamoo, "Deep scattering spectra with deep neural networks for LVCSR tasks," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2014.

[11] P. Fousek, P. L. Dognin, and V. Goel, "Evaluating deep scattering spectra with deep neural networks on large scale spontaneous speech task," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4550–4554, 2015.

[12] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 29, no. 2, pp. 254–272, 1981.

[13] F.-H. Liu, R. M. Stern, X. Huang, and A. Acero, "Efficient cepstral normalization for robust speech recognition," in *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey*, 1993.

[14] V. Joshi, N. V. Prasad, and S. Umesh, "Modified cepstral mean normalization - transforming to utterance specific non-zero mean," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2013.

[15] G. Hirsch, "Experimental Framework for the Performance Evaluation of Speech Recognition Front-ends on a Large Vocabulary Task," Ericsson, Tech. Rep., 2002.

[16] W. Fisher, G. Doddington, and K. Goudie-Marshall, "The DARPA speech recognition research database: specifications and status," in *Proc. of DARPA Workshop on Speech Recognition*, 1986.

[17] L. van der Maaten and G. Hinton, "Visualizing High-Dimensional Data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[18] J. Garofolo, D. Graff, D. Paul, and D. Pallett, "CSR-I (WSJ0) Complete LDC93S6A," *Linguistic Data Consortium*, 1993.

[19] V. Peddinti, T. N. Sainath, S. Maymon, B. Ramabhadran, D. Nahamoo, and V. Goel, "Deep Scattering Spectrum with Deep Neural Networks," in *International Conference on Acoustics, Speech, and Signal Processing*, 2014, pp. 210–214.

[20] L. J. Ba, R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[21] G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," *arXiv:1207.0580*, 2012.

[22] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on International Conference on Machine Learning*. JMLR.org, 2015, p. 448–456.

[23] Y. Qian, M. Bi, T. Tan, and K. Yu, "Very deep convolutional neural networks for noise robust speech recognition," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, 2016.

[24] J. Rownicka, P. Bell, and S. Renals, "Multi-scale octave convolutions for robust speech recognition," in *IEEE Automatic Speech Recognition and Understanding Workshop*, 2019.